PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-053400
(43)Date of publication of application : 26.02.1999

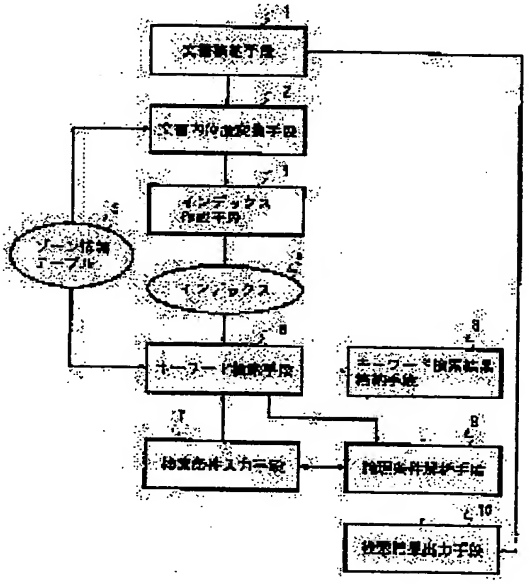
(51)Int.Cl. G06F 17/30
G06F 12/00

(21)Application number : 09-220233
(22)Date of filing : 31.07.1997

(71)Applicant : NEC CORP
(72)Inventor : AKAMINE SUSUMU

(54) STRUCTURED DOCUMENT RETRIEVAL DEVICE AND MACHINE READABLE RECORDING MEDIUM FOR RECORDING PROGRAM

(57)Abstract:
PROBLEM TO BE SOLVED: To execute zone retrieval at high speed in a structured document retrieval device.
SOLUTION: A zone information table 3 commonly defines and holds a zone name for each zone and a range of a position that the zone can occupy in the whole retrieval object document. A conversion means 2 of a position in document refers to the zone information table 3 at the time of preparing an index and prepares a zone position conversion document that moves a character string of each zone of an original document to a position indicated by the zone information table 3. Thus, it becomes possible to identify from within the document in which zone the character string exists. An index preparation means 4 makes the zone position conversion document prepared by the conversion means 2 of a position in document and prepares an index 5. A keyword retrieval means 6 executes zone retrieval only by selecting the one in which an appearance position of the keyword corresponds to a zone name of a zone instructed to be a retrieval object by a user at the time of retrieving the keyword.



LEGAL STATUS

[Date of request for examination]	31.07.1997
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	2962287
[Date of registration]	06.08.1999
[Number of appeal against examiner's decision of rejection]	

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平11－53400

(43)公開日　平成11年(1999)２月26日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ０６Ｆ	17/30		Ｇ０６Ｆ	15/419	３２０
	12/00	５４７		12/00	５４７Ｈ
				15/40	３７０Ａ

審査請求　有　　請求項の数５　ＦＤ　（全 11 頁）

(21)出願番号　特願平9－220233

(22)出願日　平成９年(1997)７月31日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　赤峯　享
東京都港区芝五丁目７番１号　日本電気株式会社内

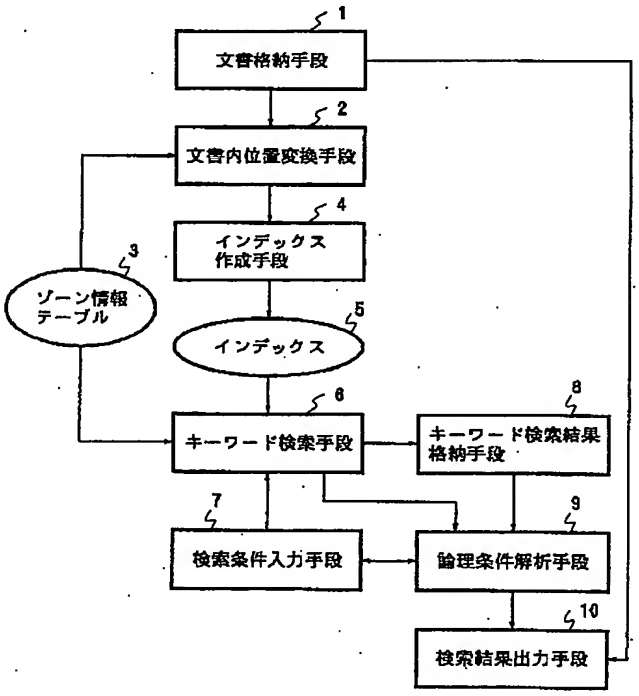
(74)代理人　弁理士　境　廣巳

(54)【発明の名称】　構造化文書検索装置及びプログラムを記録した機械読み取り可能な記録媒体

(57)【要約】

【課題】　構造化文書検索装置に於いて、ゾーン検索を高速に行う。

【解決手段】　ゾーン情報テーブル３には、各ゾーンのゾーン名と、そのゾーンが取り得る位置の範囲を全検索対象文書で共通に定めて保持している。文書内位置変換手段２は、インデックス作成時に、ゾーン情報テーブル３を参照して、オリジナルの文書の各ゾーンの文字列を、ゾーン情報テーブル３によって示される位置に移動したゾーン位置変換文書を作成する。これにより、文書内位置から文字列がどのゾーンに存在するかの識別が可能になる。インデックス作成手段４は、文書内位置変換手段２により作成されたゾーン位置変換文書を対象にしてインデックス５を作成する。キーワード検索手段６は、キーワードを検索する際に、キーワードの出現位置がユーザによって検索対象とすることが指示されたゾーンのゾーン名に対応するものだけを選択することで、ゾーン検索を行う。



1

【特許請求の範囲】

【請求項１】　複数のゾーンから構成される構造化文書が複数格納された文書格納手段と、

ゾーン位置変換文書に於ける各ゾーンの位置を示す情報が格納されたゾーン情報テーブルと、

前記文書格納手段に格納されている構造化文書中の各ゾーンを前記ゾーン情報テーブルの内容によって示される位置に移動させたゾーン位置変換文書を作成する文書内位置変換手段と、

該文書内位置変換手段によって作成されたゾーン位置変換文書に基づいて、キー文字列と、そのキー文字列が存在する構造化文書の文書識別子と、そのキー文字列のゾーン位置変換文書に於ける文書内位置とが対応して格納されたインデックスを作成するインデックス作成手段と、

検索対象にするゾーンのゾーン名とキーワードとを含む検索条件式を受け付ける検索条件入力手段と、

該検索条件入力手段が受け付けた検索条件式中のキーワードをキーにして前記インデックスを検索し、その結果得られた前記キーワードが存在する構造化文書の文書識別子，文書内位置と、前記ゾーン情報テーブルの内容とに基づいて、前記検索条件式中のゾーン名によって示されるゾーンに前記キーワードが存在する構造化文書の文書識別子を求めるキーワード検索手段とを備えたことを特徴とする構造化文書検索装置。

【請求項２】　前記ゾーン情報テーブルには、ゾーン名と、そのゾーン名のゾーンをゾーン位置変換文書内のどの位置に配置するのかを示すゾーン位置情報とが対応して格納されていることを特徴とする請求項１記載の構造化文書検索装置。

【請求項３】　前記検索条件式は、検索対象にするゾーンのゾーン名とキーワードとから構成される検索項目が論理演算記号によって複数結合された形式を有し、

前記キーワード検索手段は、前記検索条件入力手段が受け付けた検索条件式の各検索項目それぞれについて、その検索項目中のキーワードをキーにして前記インデックスを検索し、その検索結果と、前記ゾーン情報テーブルの内容とに基づいて、前記検索項目中のゾーン名によって示されるゾーンに前記キーワードが存在する構造化文書の文書識別子を求める構成を有し、且つ、

前記キーワード検索手段が求めた各検索項目毎の文書識別子と、前記検索条件入力手段が受け付けた検索条件式中の各検索項目を結合する論理演算記号とに基づいて、前記検索条件式を満足させる構造化文書の文書識別子を求める論理条件解析手段を備えたことを特徴とする請求項２記載の構造化文書検索装置。

【請求項４】　前記論理条件解析手段が求めた文書識別子の構造化文書を前記文書格納手段から読み出して出力する検索結果出力手段を備えたことを特徴とする請求項３記載の構造化文書検索装置。

2

【請求項５】　複数のゾーンから構成される構造化文書が複数格納された文書格納手段と、ゾーン位置変換文書に於ける各ゾーンの位置を示すゾーン位置情報が格納されたゾーン情報テーブルとを備えたコンピュータを、

前記文書格納手段に格納されている構造化文書中の各ゾーンを前記ゾーン情報テーブルの内容によって示される位置に移動させたゾーン位置変換文書を作成する文書内位置変換手段、

該文書内位置変換手段によって作成されたゾーン位置変換文書に基づいて、キー文字列と、そのキー文字列が存在する構造化文書の文書識別子と、そのキー文字列の文書内位置とが対応して格納されたインデックスを作成するインデックス作成手段、

検索対象にするゾーンのゾーン名とキーワードとを含む検索条件式を受け付ける検索条件入力手段、

該検索条件入力手段が受け付けた検索条件式中のキーワードをキーにして前記インデックスを検索し、その結果得られた前記キーワードを含む文書の文書識別子，文書内位置と、前記ゾーン情報テーブルの内容とに基づいて、前記検索条件式中のゾーン名によって示されるゾーンに前記キーワードが存在する構造化文書の文書識別子を求めるキーワード検索手段として機能させるためのプログラムを記録した機械読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、１件の文書が複数の論理構造（ゾーン）から構成されている構造化文書を対象にした検索技術に関し、特に、ユーザによって指定されたゾーンのみを検索対象にして構造化文書の検索を行う技術に関する。

【０００２】

【従来の技術】近年、ＳＧＭＬ(Standard Generalized Markup Language)に代表される、１件の文書を複数のゾーンで構成した構造化文書が頻繁に用いられるようになってきている。それにつれて、構造化文書の特定ゾーンのみを検索対象にした検索（ゾーン検索）を行うことが、精度の高い検索を行う上で重要な機能になってきている。

【０００３】ゾーン検索を行う従来の技術としては、例えば、検索条件を「検索対象とするゾーンの開始タグと終了タグの間にキーワードを含む文書」とし、テキスト全体を対象にして文字列照合を行うことにより、検索条件を満足させる文書を検索するようにしたものがある。しかし、この方法は、テキスト全体を対象にして文字列照合を行うので、検索時間が非常に長くなってしまう。このような問題点を解決するため、テキストの不要部分（検索対象とすることが指定されたゾーン以外のゾーン）をスキップして検索を行うようにした技術も提案されている（例えば、特開平８－１６６００号公報）。この技術によれば、文字列照合を行う範囲が少なくなるた

め、先の従来例に比較して検索時間を短くすることができる。しかし、ギガバイトクラスの大規模データに対する検索では、不要部分をスキップすることにより文字列照合範囲を例えば１０分の１程度に縮小できたとしても、百メガバイトクラスのテキストを対象にして文字列照合を行うことが必要になるため、高速な検索は望めない。

【０００４】このように、文字列照合によりゾーン検索を行う方法では、高速な検索を行うことが難しいため、大規模データに対する高速な検索を可能にするために作成された全文インデックスを利用してゾーン検索を行うことが考えられる。全文インデックスは、キー文字列がキー情報として格納されたキー情報部と、キー情報部に格納された各キー文字列それぞれについてそのキー文字列が存在する文書の文書識別子，文書内位置が位置情報として格納された位置情報部とから構成されるものであり、このような全文インデックスを利用してゾーン検索を行う方法としては、下記（Ａ）～（Ｃ）の３つの方法が考えられる。

【０００５】

（Ａ）ゾーンに関する情報をキー情報部に持たせる方法。

（Ｂ）ゾーンに関する情報を位置情報部に持たせる方法。

（Ｃ）全文インデックスとは別のゾーンに関するインデックスを作成する方法。

【０００６】（Ａ）の方法では、全文インデックスのキー情報部に、キー文字列とそのキー文字列が存在するゾーンのゾーン名とのペアからなるキー情報を格納しておく。その際、複数のゾーンに存在するキー文字列については、各ゾーン毎にキー文字列とゾーン名との対からなるキー情報を格納する。位置情報部には、各キーワード毎に該当する文書の文書識別子，文書内位置が格納される。そして、検索時には、ユーザによって指定されたゾーン名とキーワードとをキーにして全文インデックスを検索することにより、上記ゾーン名のゾーンに、上記キーワードを含む文書を探し出すようにしている。

【０００７】（Ｂ）の方法では、位置情報部に格納する位置情報に、位置情報として文書識別子，文書内位置の他にゾーン名も併せ持たせておく。そして、検索時には、先ず、ユーザによって指定されたキーワードをキーにして全文インデックスを検索することにより、上記キーワードを含む文書の位置情報を全て求め、その後、上記位置情報の中からユーザによって指定されたゾーン名を含む位置情報を選択することにより、ゾーン検索を行うようにしている。

【０００８】（Ｃ）の方法では、全文インデックスとは別に、検索対象とする全ての文書それぞれの、各ゾーンの開始位置，終了位置が格納されたゾーン用インデックスを作成しておく。そして、検索時には、先ず、全文インデックスを検索することにより、ユーザが指定したキーワードを含む文書の位置情報を取得する。その後、ゾーン用インデックスを検索し、上記文書の、ユーザによって指定されたゾーンの開始位置，終了位置を取得する。更に、位置情報中の文書内位置と取得した上記ゾーンの開始位置，終了位置とに基づいて、上記文書のユーザによって指定されたゾーン内に、ユーザによって指定されたキーワードが存在するか否かをチェックすることにより、ゾーン検索を行う（例えば、特開平８－３１４９６６号公報）。

【０００９】

【発明が解決しようとする課題】しかしながら、上述した（Ａ）の方法は、ゾーンの数に比例してキー情報数が多くなるため、全文インデックスのサイズが大きくなるという問題がある。更に、全文インデックスのサイズが大きくなることにより、検索速度が低下するという問題もある。

【００１０】また、（Ｂ）の方法は、全文インデックスの位置情報部の容量が多くなるため、全文インデックスのサイズが大きくなるという問題がある。更に、位置情報部からの読み出し量が多くなるため、検索速度が低下するという問題もある。

【００１１】また、（Ｃ）の方法は、検索対象とする全ての文書それぞれの、各ゾーンの開始位置，終了位置を格納したゾーン用インデックスが必要になるため、インデックスサイズが大きくなるという問題がある。更に、全文インデックスを検索することにより探し出した各該当文書について、ゾーン用インデックスを検索し、ユーザが指定したキーワードが、ユーザが指定したゾーン内に存在するか否かをチェックする必要があるため、この処理がオーバーヘッドになって検索速度が低下してしまうという問題が生じる。

【００１２】そこで、本発明の目的は、全文インデックスを利用したゾーン検索に於いて、インデックスサイズを小さくし、且つ検索速度を高速化することにある。

【００１３】

【課題を解決するための手段】本発明の構造化文書検索装置は、上記目的を達成するため、複数のゾーンから構成される構造化文書が複数格納された文書格納手段と、ゾーン位置変換文書に於ける各ゾーンの位置を示す情報が格納されたゾーン情報テーブルと、前記文書格納手段に格納されている構造化文書中の各ゾーンを前記ゾーン情報テーブルの内容によって示される位置に移動させたゾーン位置変換文書を作成する文書内位置変換手段と、該文書内位置変換手段によって作成されたゾーン位置変換文書に基づいて、キー文字列と、そのキー文字列が存在する構造化文書の文書識別子と、そのキー文字列のゾーン位置変換文書に於ける文書内位置とが対応して格納されたインデックスを作成するインデックス作成手段と、検索対象にするゾーンのゾーン名とキーワードとを

含む検索条件式を受け付ける検索条件入力手段と、該検索条件入力手段が受け付けた検索条件式中のキーワードをキーにして前記インデックスを検索し、その結果得られた前記キーワードが存在する構造化文書の文書識別子，文書内位置と前記ゾーン情報テーブルの内容とに基づいて、前記検索条件式中のゾーン名によって示されるゾーンに前記キーワードが存在する構造化文書の文書識別子を求めるキーワード検索手段とを備えたものである。

【００１４】この構成に於いては、ゾーン検索を行うための準備として、文書内位置変換手段が、文書格納手段に格納されている構造化文書中の各ゾーンをゾーン情報テーブルの内容によって示される位置に移動させたゾーン位置変換文書を作成し、インデックス作成手段が、文書内位置変換手段によって作成されたゾーン位置変換文書に基づいて、キー文字列と、そのキー文字列が存在する構造化文書の文書識別子と、そのキー文字列のゾーン位置変換文書に於ける文書内位置とが対応して格納されたインデックスを作成する。

【００１５】そして、ゾーン検索時に、ユーザが検索対象にするゾーンのゾーン名とキーワードとを含む検索条件式を入力すると、検索条件入力手段がそれを受け付け、キーワード検索手段が上記検索条件式中のキーワードをキーにしてインデックスを検索し、その結果得られた前記キーワードが存在する構造化文書の文書識別子，文書内位置とゾーン情報テーブルの内容とに基づいて、上記検索条件式中のゾーン名によって示されるゾーンに前記キーワードが存在する構造化文書の文書識別子を求める。

【００１６】

【発明の実施の形態】次に本発明の実施の形態について図面を参照して詳細に説明する。

【００１７】図１は本発明の実施例のブロック図であり、文書格納手段１と、文書内位置変換手段２と、ゾーン情報テーブル３と、インデックス作成手段４と、インデックス５と、キーワード検索手段６と、検索条件入力手段７と、キーワード検索結果格納手段８と、論理条件解析手段９と、検索結果出力手段１０とから構成されている。

【００１８】文書格納手段１には、検索対象となる複数の構造化文書が格納されている。図２は文書格納手段１の内容例を示した図である。本実施例では、説明を簡単にするため、文書格納手段１には、文書識別子ＩＤ１，ＩＤ２の２つの構造化文書２１，２２が格納されているとする。各構造化文書２１，２２は、それぞれゾーン名「全体」，「発明の名称」，「要約」，「目的」，「構成」の各ゾーンから構成され、各ゾーンは、ゾーン開始タグ，ゾーン終了タグ等により分割されている。

【００１９】ゾーン情報テーブル３には、文書内位置変換手段２で作成するゾーン位置変換文書に於ける各ゾーンの位置を示す情報が格納されている。図３はゾーン情報テーブル３の内容例を示した図であり、ゾーン名と、そのゾーン名のゾーンをゾーン位置変換文書内のどの位置に配置するのかを示すゾーン位置情報とが格納されている。図３の例は、ゾーン名「全体」，「発明の名称」，「要約」，「目的」，「構成」のゾーンを、それぞれゾーン位置変換文書内の「第１文字目～第２０００文字目」，「第１文字目～第５００文字目」，「第５０１文字目～第２０００文字目」，「第５０１文字目～第１０００文字目」，「第１００１文字目～第２０００文字目」に配置することを示している。

【００２０】文書内位置変換手段２は、ゾーン情報テーブル３を参照し、文書格納手段１に格納されている構造化文書に対して、各ゾーンの文字列をゾーン情報テーブル３中のゾーン位置情報によって示される位置に移動したゾーン位置変換文書を作成する機能を有する。従って、各構造化文書では異なる位置に存在していた各ゾーンの文字列は、文書内位置変換手段２によって作成されたゾーン位置変換文書では、同じ範囲に存在することになる。

【００２１】インデックス作成手段４は、文書内位置変換手段２で作成された各構造化文書に対応するゾーン位置変換文書に基づいてインデックス５を作成する機能を有する。インデックス５は、図４に示すように、キー情報部５１と、位置情報部５２とを有している。キー情報部５１にはＮ文字組や単語等のキー情報が格納され、位置情報部５１には、キー情報が存在する構造化文書の文書識別子と、そのキー情報が存在するゾーン位置変換文書内の位置とが格納される。ここで、位置情報「ｉ－ｊ」は、文書識別子ＩＤｉの構造化文書に対応するゾーン位置変換文書の第ｊ文字目を表している。従って、図４の例の第１番目のエリアは、文字「文」が文書識別子ＩＤ１の構造化文書２１中に存在し、それと対応するゾーン位置変換文書２１’では第１文字目，第５０４文字目に存在することと、文字「文」が文書識別子ＩＤ２の構造化文書２２中に存在し、それと対応するゾーン位置変換文書２２’では第１文字目，第５０１文字目に存在することを表している。

【００２２】検索条件入力手段７は、ユーザによって入力された検索条件式を受け付ける機能，論理条件解析手段９を利用して検索条件式を検索項目に分解する機能，検索項目をキーワード検索手段６に渡す機能等を有する。ユーザが入力する検索条件式は、検索対象とするゾーン名とキーワードとのペアからなる検索項目を１つ或いは複数含むものであり、検索項目を複数含む場合は、各検索項目は、ＡＮＤ，ＯＲ等の論理演算記号によって結合される。図５は、ユーザが入力する検索条件式の１例を示した図であり、２つの検索項目が論理演算記号ＡＮＤによって結合されている。この検索条件式は、ゾーン名「発明の名称」のゾーンにキーワード「検索」を含

み、且つゾーン名「要約」のゾーンにキーワード「インデックス」を含む構造化文書の検索を指示するものである。

【００２３】キーワード検索手段６は、検索条件入力手段７から渡された各検索項目中のキーワードをキーにしてインデックス５を検索することにより、各検索項目それぞれについて、その検索項目中のキーワードが現れる文書の文書識別子，文書内位置を全て求める機能や、各検索項目それぞれについて、ゾーン情報テーブル３を参照して検索項目に含まれているゾーン名によって示されるゾーンのゾーン位置を求める機能や、各検索項目それぞれについて、その検索結果（文書識別子，文書内位置）の中に上記ゾーン位置内の位置を示す検索結果があれば、その検索結果中の文書識別子とそれが何番目の検索項目についてのものなのかを示す情報とをペアにしてキーワード検索結果格納手段８に格納する機能等を有する。

【００２４】論理条件解析手段９は、キーワード検索結果格納手段８に格納されている検索項目毎の検索結果（文書識別子）と、検索条件入力手段７が受け付けた検索条件式中の各検索項目を結合する論理演算記号とに基づいて、上記検索条件式を満足させる構造化文書の文書識別子を求める機能を有する。

【００２５】検索結果出力手段１０は、論理条件解析手段９が求めた文書識別子を有する構造化文書を文書格納手段１から取り出し、プリンタ，ＣＲＴ等の出力装置（図示せず）に出力する機能を有する。

【００２６】図６は文書内位置変換手段２の処理例を示す流れ図、図７はインデックス作成手段４の処理例を示す流れ図、図８は検索条件入力手段７の処理例を示す流れ図、図９は検索条件入力手段７から検索条件式が渡されたときの論理条件解析手段９の処理例を示す流れ図、図１０はキーワード検索手段６から終了通知が送られてきたときの論理条件解析手段９の処理例を示す流れ図、図１１はキーワード検索手段６の処理例を示す流れ図であり、以下各図を参照して本実施例の動作を説明する。

【００２７】先ず、インデックス５の作成時の動作を説明する。

【００２８】インデックス５の作成時、文書内位置変換手段２は、図６の流れ図に示すように、文書格納手段１から未処理の構造化文書を１つ入力する（Ｓ６１）。その後、文書内位置変換手段２は、入力した構造化文書の各ゾーン中の文字列をゾーン情報テーブル３中のゾーン位置情報によって示される位置に移動させたゾーン位置変換文書を作成し（Ｓ６３）、作成したゾーン位置変換文書とＳ６１で入力した構造化文書の文書識別子とをインデックス作成手段４に渡す（Ｓ６４）。以上の処理を未処理の構造化文書がなくなるまで（Ｓ６２がＮＯ）、繰り返し行う。

【００２９】本実施例の場合、文書格納手段１には図２に示すような文書識別子ＩＤ１，ＩＤ２の構造化文書２１，２２が格納され、更に、ゾーン情報テーブル３の内容は図３に示すものになっているので、文書内位置変換手段２は、図１２に示すようなゾーン位置変換文書２１’，２２’を順次作成してインデックス作成手段４に渡すことになる。

【００３０】つまり、文書内位置変換手段２は、図３に示したゾーン情報テーブル３の内容を参照し、その内容に従って、構造化文書２１のゾーン「発明の名称」中に存在する文字列「文書検索装置」を第１文字目から始まる位置に移動させ、ゾーン「目的」中に存在する文字列「高速に文書を検索する。」を第５０１文字目から始まる位置に移動させ、ゾーン「構成」中に存在する文字列「インデックス作成手段と……。」を第１００１文字目から始まる位置に移動させた図１２に示すようなゾーン位置変換文書２１’を作成して文書識別子ＩＤ１と共にインデックス作成手段４に渡す。同様に、文書内位置変換手段２は、構造化文書２２のゾーン「発明の名称」中に存在する文字列「文書処理装置」を第１文字目から始まる位置に移動させ、ゾーン「目的」中に存在する文字列「文書を……。」を第５０１文字目から始まる位置に移動させ、ゾーン「構成」中に存在する文字列「検索手段と……。」を第１００１文字目から始まる位置に移動させた図１２に示すようなゾーン位置変換文書２２’を作成して文書識別子ＩＤ２と共にインデックス作成手段４に渡す。

【００３１】図１２から判るように、変換処理後の各ゾーン位置変換文書２１’，２２’は、ゾーン名「全体」，「発明の名称」，「要約」，「目的」，「構成」の各ゾーンが、ゾーン情報テーブル３中のゾーン位置情報によって示される位置に必ず存在することになる。

【００３２】インデックス作成手段４は、文書内位置変換手段２からゾーン位置変換文書，文書識別子が渡されると、図７の流れ図に示すように、ゾーン位置変換文書の先頭位置に注目する（Ｓ７１）。そして注目位置に、インデックス５のキー情報部５１に格納すべき文字が存在するか否かを判断する（Ｓ７２）。格納すべき文字か否かの判断は、例えば、空白文字，句読点等、格納する必要のない文字を予め定めておき、注目位置に存在する文字がそれ以外の文字であるか否かを判断することにより行う。

【００３３】そして、Ｓ７２に於いて、格納すべき文字が注目位置に存在すると判断した場合（Ｓ７２がＹＥＳ）は、その文字が既に格納済みか否かを判断する（Ｓ７３）。格納済みでないと判断した場合は、注目位置に存在する文字をインデックス５のキー情報部５１に格納すると共に、文書内位置変換手段２から渡された文書識別子と文書内位置（現在の注目位置）とからなる位置情報を位置情報部５２に格納する（Ｓ７３がＮＯ，Ｓ７４）。これに対して、格納済みであると判断した場合

は、位置情報部５２に文書内位置変換手段２から渡された文書識別子と文書内位置とからなる位置情報を位置情報部５２に格納する（Ｓ７３がＹＥＳ，Ｓ７５）。

【００３４】Ｓ７４，７５の処理が終了すると、インデックス作成手段４は、注目位置を次の位置に移し（Ｓ７６）、前述したと同様の処理を行う。また、Ｓ７２で格納すべき文字が注目位置に存在しないと判断した場合も、Ｓ７６の処理を行う。

【００３５】以上の処理を文書内位置変換手段２から渡されたゾーン位置変換文書の終わりまで（Ｓ７７がＹＥＳ）、繰り返し行う。

【００３６】本実施例の場合、インデックス作成手段４には、図１２に示すようなゾーン位置変換文書２１’，文書識別子ＩＤ１と、ゾーン位置変換文書２２’，文書識別子ＩＤ２とが渡されるので、インデックス作成手段４に於いては、次のような処理が行われることになる。

【００３７】文書内位置変換手段２からゾーン位置変換文書２１’と文書識別子ＩＤ１とが渡された場合は、インデックス作成手段４は、先頭位置に注目したときに、文字「文」をキー情報部５１に格納し、位置情報「１－１」を位置情報部５２に格納する（Ｓ７１，Ｓ７４）。また、インデックス作成手段４は、ゾーン位置変換文書２１’中の次に位置（第２文字目）に注目したときは、注目位置に存在する文字「書」をキー情報部５１に格納し、位置情報「１－２」を位置情報部５２に格納する（Ｓ７４）。また、例えば、注目位置をゾーン位置変換文書２１’の第５０４文字目にしたときは、注目位置に存在する文字「文」は既に格納済みであるので、位置情報部５２中の上記文字「文」に対応するエントリに位置情報「１－５０４」を格納することになる（Ｓ７５）。このような処理を、ゾーン位置変換文書２１’の終わりまで行う。ゾーン位置変換文書２２’と文書ＩＤとが渡された場合も、インデックス作成手段４は前述したと同様の処理を行う。この結果、インデックス５の内容は、図４に示すものとなる。

【００３８】尚、ここでは、説明を簡単に行うため、キー情報部５１に格納する文字列の文字長を１文字としたが、これに限られるものではなく、文字長が２以上のＮ文字組でも、単語であっても構わない。

【００３９】次に、ゾーン検索時の動作について説明する。

【００４０】ゾーン検索を行う場合、ユーザは、検索対象とするゾーンのゾーン名とキーワードとのペアからなる検索項目を１つ或いは複数含む検索条件式を検索条件入力手段７に入力する。前述したように、検索項目を複数含む検索条件式の場合は、各検索項目は、ＡＮＤ，ＯＲ等の論理演算記号によって結合されている。

【００４１】今、例えば、ユーザが検索条件式として、図５に示した検索条件式「（発明の名称＝検索）ＡＮＤ（要約＝インデックス）」を検索条件入力手段７に入力

したとする。この検索条件式は、前述したように、ゾーン名「発明の名称」の部分に文字列「検索」が現れ、且つゾーン名「要約」の部分に文字列「インデックス」が現れる構造化文書の検索を指示するものである。

【００４２】ユーザが検索条件式「（発明の名称＝検索）ＡＮＤ（要約＝インデックス）」を入力すると、検索条件入力手段７は、図８の流れ図に示すように、それを受け付け、論理条件解析手段９に渡す（Ｓ８１，Ｓ８２）。

【００４３】論理条件解析手段９は、検索条件式「（発明の名称＝検索）ＡＮＤ（要約＝インデックス）」が渡されると、図９の流れ図に示すように、検索条件式を第１番目の検索項目「発明の名称＝検索」と、第２番目の検索項目「要約＝インデックス」との２つの検索項目に分割し、それらを検索条件入力手段７に返す（Ｓ９１，Ｓ９２）。

【００４４】検索条件入力手段７は、論理条件解析手段９から第１番目，第２番目の検索項目「発明の名称＝検索」，「要約＝インデックス」を受け取ると、それらをキーワード検索手段６に渡す（図８のＳ８３，Ｓ８４）。

【００４５】キーワード検索手段６は、検索条件入力手段７から第１番目，第２番目の検索項目「発明の名称＝検索」，「要約＝インデックス」が渡されると、図１１の流れ図に示すように、その内の１つに注目する（Ｓ１１１）。

【００４６】今、例えば、第１番目の検索項目「発明の名称＝検索」に注目したとすると、キーワード検索手段６は、先ず、第１番目の検索項目「発明の名称＝検索」中のキーワード「検索」をキーにしてインデックス５を検索することにより、キーワード「検索」が現れるゾーン位置変換文書の文書識別子と、文書内位置とを求める（Ｓ１１３）。本実施例の場合、インデックス５の内容は、図５に示すものになっているので、Ｓ１１３を行うことにより、キーワード「検索」が、文書識別子ＩＤ１のゾーン位置変換文書２１’の第３文字目～第４文字目と、文書識別子がＩＤ２のゾーン位置変換文書２２’の第１００１文字目～第１００２文字目に現れることが求められる。

【００４７】その後、キーワード検索手段６は、ゾーン情報テーブル３を参照し、第１番目の検索項目中のゾーン名「発明の名称」によって示されるゾーンのゾーン位置を求める（Ｓ１１４）。本実施例の場合、「第１文字目～第５００文字目」がゾーン位置として求められる（図３参照）。

【００４８】次いで、キーワード検索手段６は、Ｓ１１３の検索結果の中に、Ｓ１１４で求めたゾーン位置内の位置を示すものがあれば、その検索結果中の文書識別子とそれが第１番目の検索項目についてのものであることを示す情報とをペアにしてキーワード検索結果格納手段

８に格納する（Ｓ１１５）。この例では、Ｓ１１３の検索結果が、「文書識別子ＩＤ１のゾーン位置変換文書２１'の第３文字目～第４文字目」，「文書識別子がＩＤ２のゾーン位置変換文書２２'の第１００１文字目～第１００２文字目」で、Ｓ１１４で求めたゾーン位置が「第１文字目～第５００文字目」であるので、キーワード検索手段６は、文書識別子ＩＤ１とそれが第１番目の検索項目についてのものであることを示す情報とをペアにしてキーワード検索結果格納手段８に格納することになる。

【００４９】その後、キーワード検索手段６は、第２番目の検索項目「要約＝インデックス」に注目し（Ｓ１１１）、前述したと同様の処理を行う（Ｓ１１３～Ｓ１１５）。第２番目の検索項目の場合、キーワード「インデックス」は、文書識別子ＩＤ１のゾーン位置変換文書２１'の第１００１文字目～第１００６文字目に現れ、ゾーン名「要約」のゾーンのゾーン位置は、「第５０１文字目～第２０００文字目」であるので、キーワード検索手段６は、文書識別子ＩＤ１とそれが第２番目の検索項目についてのものであることを示す情報とをペアにしてキーワード検索結果格納手段８に格納する。

【００５０】そして、検索条件入力手段７から渡された全ての検索項目について上述した処理を行うと（Ｓ１１２がＮＯ）、キーワード検索手段６は、論理条件解析手段９に対して終了通知を送る（Ｓ１１６）。

【００５１】論理条件解析手段９は、キーワード検索手段６から終了通知が送られてくると、図１０の流れ図に示すように、キーワード検索結果格納手段８に格納されている各検索項目についての検索結果と、検索条件入力手段７から渡された検索条件式中の論理演算記号とに基づいて、検索条件式を満足させる構造化文書の文書識別子を求め、それを検索結果出力手段１０に渡す（Ｓ１０１，Ｓ１０２）。

【００５２】この例の場合、キーワード検索結果格納手段８には、第１番目，第２番目の検索項目の検索結果としてそれぞれ文書識別子「ＩＤ１」，「ＩＤ１」が格納され、検索条件式中の第１番目の検索項目と第２番目の検索項目とを結合する論理演算式が「ＡＮＤ」であることから、両方の検索結果中に存在する文書識別子「ＩＤ１」を検索結果出力手段１０に渡すことになる。

【００５３】検索結果出力手段１０は、文書識別子「ＩＤ１」が渡されると、文書格納手段１から文書識別子が「ＩＤ１」の構造化文書２１を読み込み、プリンタ，ＣＲＴ等の出力装置（図示せず）に出力する。

【００５４】図１３は、図１に示した構造化文書検索装置のハードウェア構成の一例を示したブロック図であり、コンピュータ１３１と、記録媒体１３２と、記憶装置１３３とから構成されている。記録媒体１３２は、磁気ディスク，半導体メモリ，その他の記録媒体であり、コンピュータ１３１を構造化文書検索装置として機能させるためのプログラムが記録されている。

【００５５】記録媒体１３２に記録されているプログラムは、コンピュータ１３１によって読み込まれ、コンピュータ１３１の動作を制御することにより、コンピュータ１３１上に図１に示した文書内位置変換手段２，インデックス作成手段４，キーワード検索手段６，検索条件入力手段７，論理条件解析手段９，検索結果出力手段１０を実現する。尚、文書格納手段１，ゾーン情報テーブル３，インデックス５，キーワード検索結果格納手段８は、記憶装置１３３上に構成される。

【００５６】

【発明の効果】以上説明したように、本発明の構造化文書検索装置によれば、従来の全文インデックスを利用してゾーン検索を行う従来の技術に比較してインデックスサイズを小さくすることができ、且つ検索速度を高速化することができる。その理由は、ゾーンの位置情報が全ての構造化文書で共通になるような形でインデックスを作成するため、検索時に、非常に小規模なゾーン情報テーブルを参照するだけでゾーン検索を行うことができるからである。

【図面の簡単な説明】

【図１】本発明の実施例のブロック図である。

【図２】文書格納手段１の内容例を示す図である。

【図３】ゾーン情報テーブル３の内容例を示す図である。

【図４】インデックス５の内容例を示す図である。

【図５】検索条件式の一例を示す図である。

【図６】文書内位置変換手段２の処理例を示す流れ図である。

【図７】インデックス作成手段４の処理例を示す流れ図である。

【図８】検索条件入力手段７の処理例を示す流れ図である。

【図９】検索条件入力手段７から検索条件式が渡されたときの論理条件解析手段９の処理例を示す流れ図である。

【図１０】キーワード検索手段６から終了通知が送られてきたときの論理条件解析手段９の処理例を示す流れ図である。

【図１１】キーワード検索手段６の処理例を示す流れ図である。

【図１２】文書内位置変換手段２で作成されたゾーン位置変換文書の一例を示す図である。

【図１３】図１に示した構造化文書検索装置のハードウェア構成の一例を示すブロック図である。

【符号の説明】

１…文書格納手段

２…文書内位置変換手段

３…ゾーン情報テーブル

４…インデックス作成手段

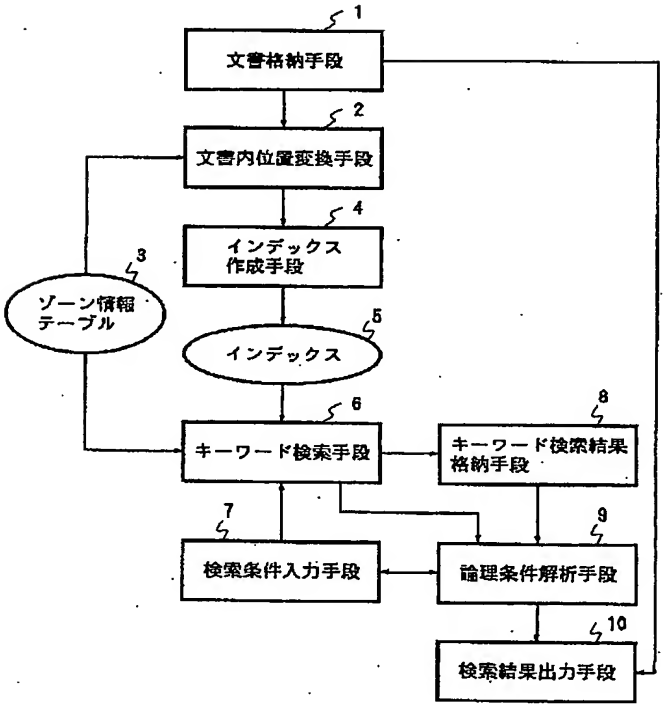
13

５…インデックス
６…キーワード検索手段
７…検索条件入力手段
８…キーワード検索結果格納手段
９…論理条件解析手段
１０…検索結果出力手段
２１，２２…構造化文書

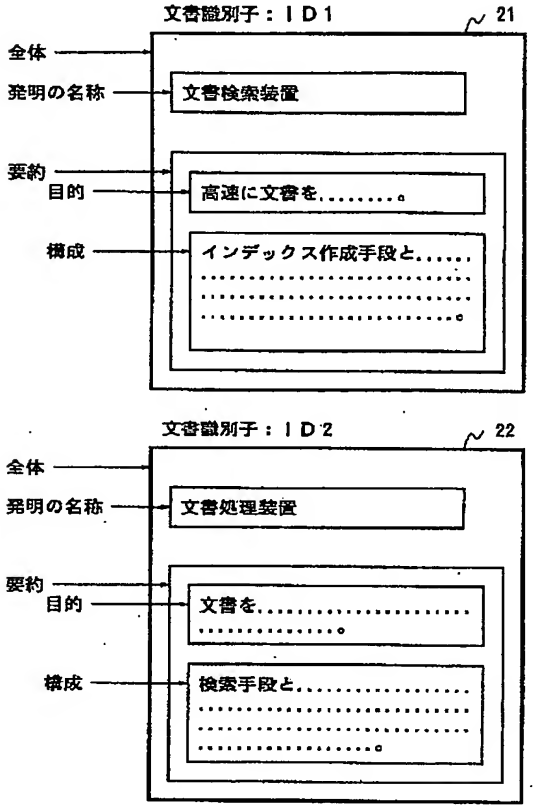
14

２１'，２２'…ゾーン位置変換文書
５１…キー情報部
５２…位置情報部
１３１…コンピュータ
１３２…記録媒体
１３３…記憶装置

【図１】



【図２】



【図３】

3 ゾーン情報テーブル

ゾーン名	ゾーン位置情報
全体	1文字目 ～ 2000文字目
発明の名称	1文字目 ～ 500文字目
要約	501文字目 ～ 2000文字目
目的	501文字目 ～ 1000文字目
構成	1001文字目 ～ 2000文字目

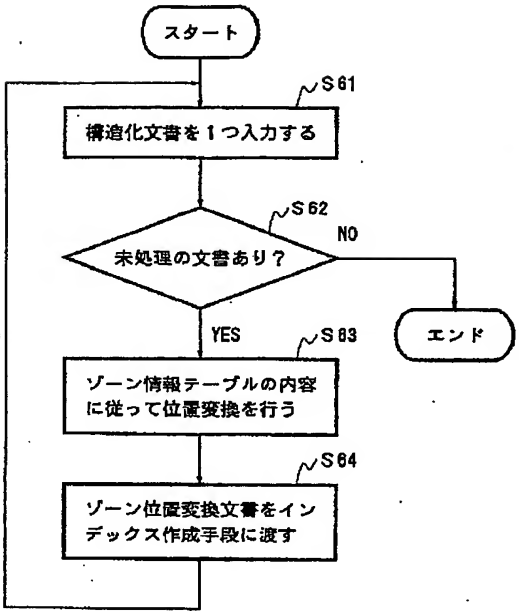
【図５】

【図４】

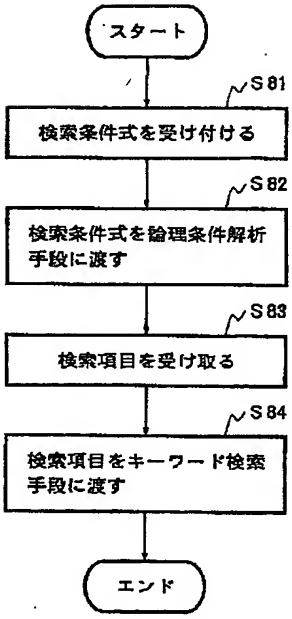
5 インデックス

51 キー情報	52 位置情報（文書識別－文書内位置）
文	1-1, 1-504, 2-1, 2-501
書	1-2, 1-505, 2-2, 2-502
検	1-3, 2-1001
索	1-4, 2-1002
装	1-5, 2-5
置	1-6, 2-6
高	1-501
速	1-502
・・・	・・・
イ	1-1001
ン	1-1002
デ	1-1003
ッ	1-1004
ク	1-1005
ス	1-1006
・・・	・・・

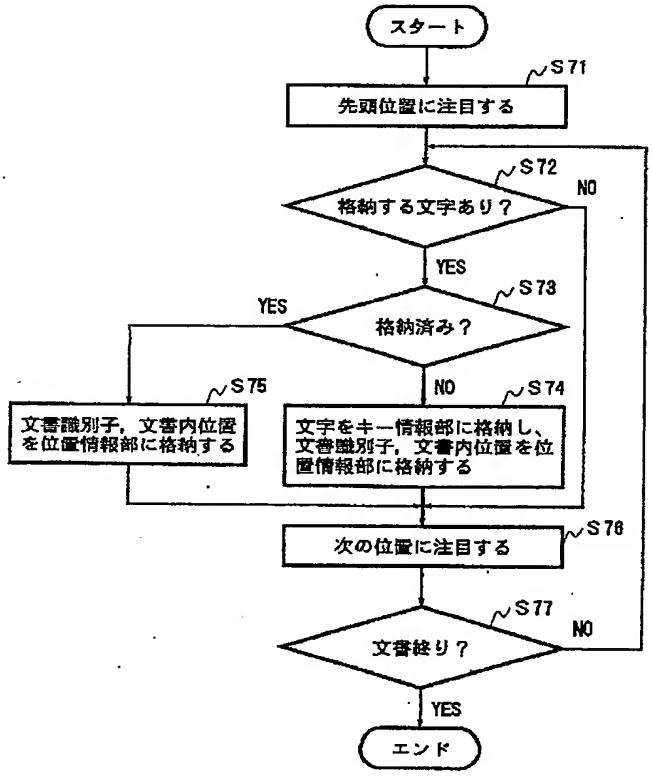
【図６】



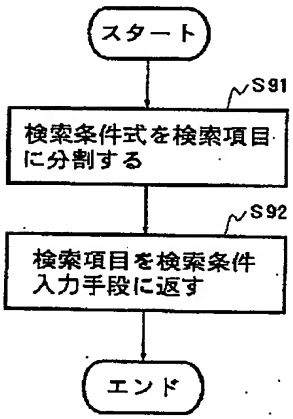
【図８】



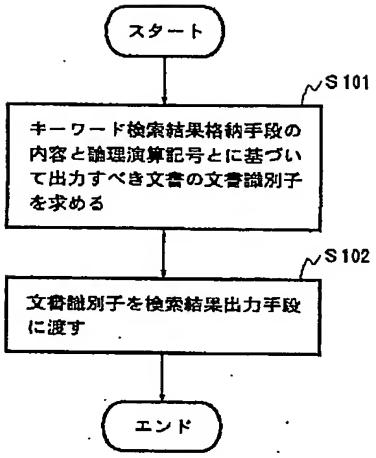
【図７】



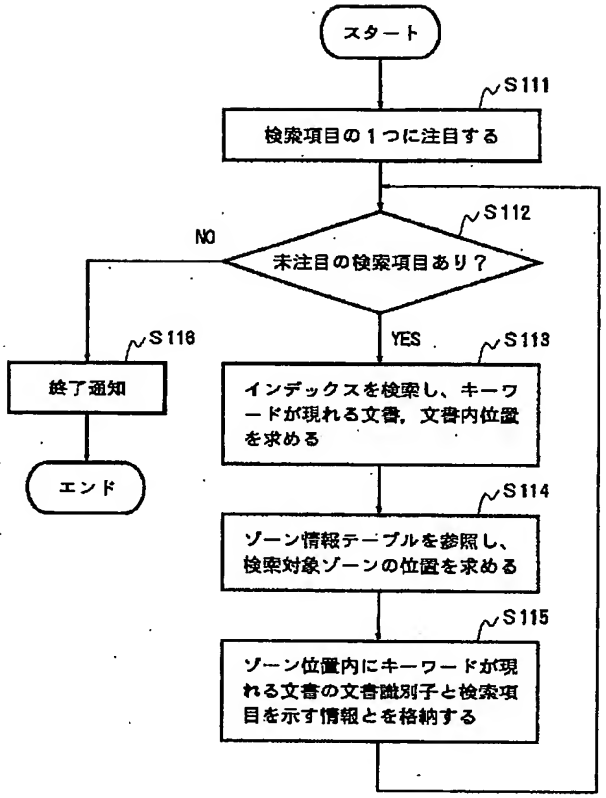
【図９】



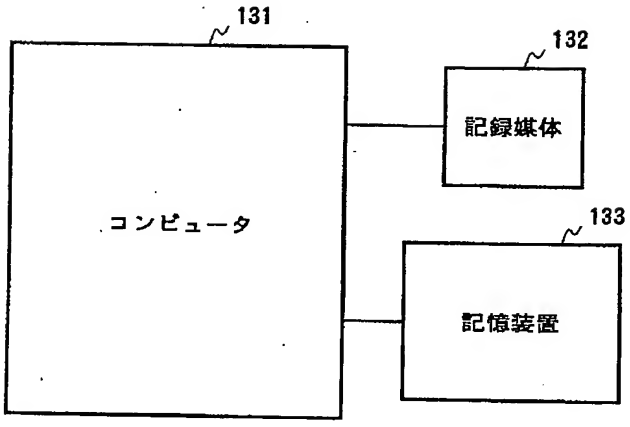
【図１０】


スタート
S101
キーワード検索結果格納手段の内容と論理演算記号とに基づいて出力すべき文書の文書識別子を求める
S102
文書識別子を検索結果出力手段に渡す
エンド

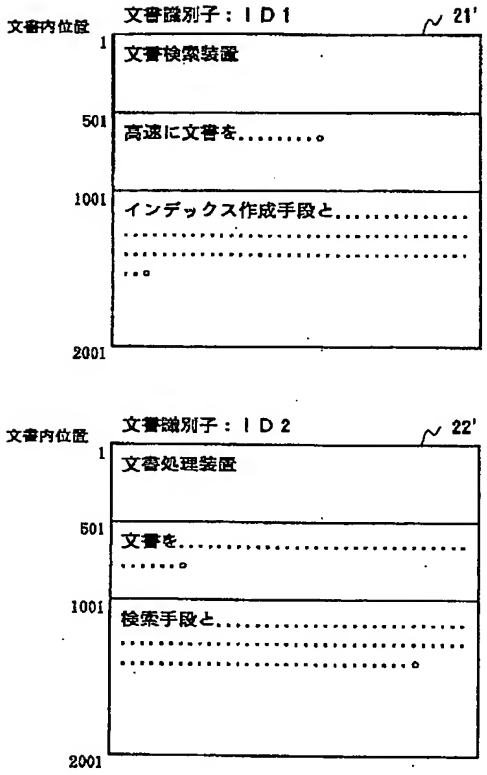
【図１１】


スタート
S111
検索項目の１つに注目する
S112
NO
未注目の検索項目あり？
YES
S116
終了通知
エンド
S113
インデックスを検索し、キーワードが現れる文書，文書内位置を求める
S114
ゾーン情報テーブルを参照し、検索対象ゾーンの位置を求める
S115
ゾーン位置内にキーワードが現れる文書の文書識別子と検索項目を示す情報とを格納する

【図１３】


131
コンピュータ
132
記録媒体
133
記憶装置

【図１２】


文書内位置
文書識別子：ＩＤ１
21'
1
文書検索装置
501
高速に文書を.....。
1001
インデックス作成手段と.....
.....。
2001
文書内位置
文書識別子：ＩＤ２
22'
1
文書処理装置
501
文書を.....
.....。
1001
検索手段と.....
.....。
2001